

Evaluation of the usage of probabilistic context-free grammars for the detection of binding sites within protein sequences

Witold Dyrka and Jean-Christophe Nebel

Faculty of Computing, Information Systems & Mathematics, Kingston University
Kingston-upon-Thames, KT1 2EE, UK

J.Nebel@kingston.ac.uk

Abstract

The analysis of a protein, through the evaluation of interactions between the amino acids composing its sequence, is a very challenging problem where pattern recognition techniques based on Hidden Markov Models (HMMs) have proved to be the most efficient. However, HMMs have limitations. According to formal language theory, their expressive power is similar to Probabilistic Regular Grammars (PRGs). A more powerful grammar, Probabilistic Context-Free Grammar (PCFG), has been applied successfully for the prediction of RNA structure. However, its utilisation in protein pattern recognition is a more challenging task due to the larger amino acid alphabet and less straightforward relations between residues. We developed a PCFG which uses quantitative properties of amino acids in order to reduce the number of symbols. Our PCFGs proved their ability to detect binding sites with high accuracy. Moreover, they show good potential for detecting sites which cannot be described by a single pattern.

Rational

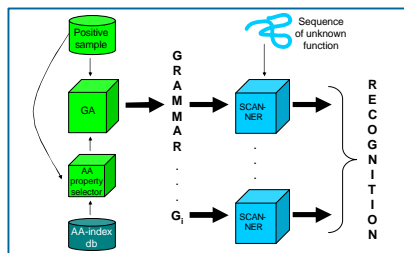
- Detection of binding sites in a protein sequence is key to the problem of protein annotation
- State of the art Hidden Markov Models (HMMs) based methods [1] cannot take into account long distance relationships between amino acids (AA)
- Probabilistic Context-Free Grammar (PCFG) is a more powerful language which has been successfully applied to the prediction of RNA secondary structures [1,2]

Method

Principle

Use quantitative properties of AAs to limit the number of symbols present in the PCFGs describing the binding sites of interest. Grammars are generated using a Genetic Algorithm (GA).

General scheme



- Selection of AA property relevant to binding site
- Extraction of grammar rules using GA from a positive training set
- Scanning of sequence of interest using grammar
- Detection of binding site if probability of a position is above an automatically generated threshold
- Grammars based upon different properties are combined to achieve more robust results
- Amino acid properties
 - Quantitative properties of AA:
 - AAindex[3]: over 500 properties, e.g. charge, accessibility and volume
 - MSDsite[4]: ligand propensities
 - For a given property, 3 non-terminal symbols (NT) created for low, med. and high level of the property
 - Association of each AA with each NT is calculated on the basis of normalised value of the property
 - Finally, these values of association are processed, so the sum of probabilities equals 1 for each NT

Implementation

CKY based PCFG parser feat. standardised score for window size independent scanning. Genetic Algorithm based on M. Wall's GALib 2.4.6

Are PCFGs a practical tool?

- YES, PCFGs achieved high *Precision* and *Recall* rates for all tested patterns

Overall performance of sequence annotation:

Pattern	Max F1	Precision	Recall	Rc Pr=1
PS00219	1.00	1.00	1.00	1.00
PS00063	0.89	0.97	0.81	0.42
PS00307	0.91	1.00	0.84	0.84
Zn-finger*	0.81	0.75	0.87	0.22

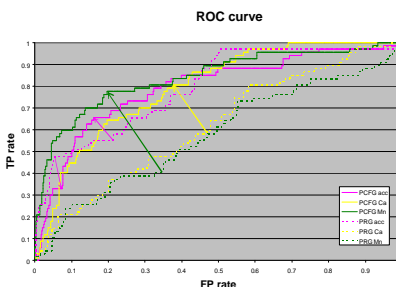
Effect of property type and combination in annotation and site localisation for PS00307:

Property	Annotation		Localisation	
	Max F1	Rc Pr=1	Exact	50%
accessibility +Ca+Mn	0.91	0.84	0.75	1.00
accessibility	0.61	0.00	0.76	0.99
Ca	0.44	0.06	0.45	0.85
Mn	0.52	0.00	0.37	0.88

- Addition of secondary structure information of binding site and its neighbourhood can improve results, particularly when *Precision* is low.

Does context-free bring added value?

- Comparison of performance of regular and context-free grammars for some properties of PS00307 pattern:



- YES, some features of patterns cannot be properly represented in the terms of regular grammars

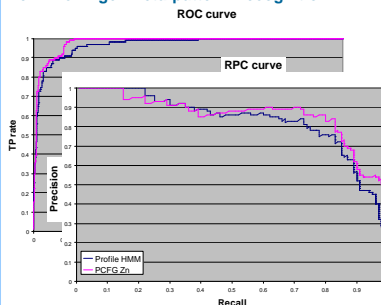
- e.g. ligand propensities (Ca, Mn...)

- Context-free does not matter when properties depends more on amino acid nature than interaction with others

- e.g. accessibility

Can we match state-of-the-art?

- YES, PCFGs slightly outperformed Profile HMMs for Zinc-finger meta-pattern* recognition



- However, due to reduced information, PROSITE and Profile HMMs were usually better with simple patterns

Conclusions

- PCFG based on quantitative representation of AA properties proved to be successful for a range of protein patterns
- Combining grammars based on different properties proved to be efficient strategy
- Secondary structure grammars improved significantly results for some patterns
- Some sequence features appear to require context-free description
- Results of PCFGs for meta-pattern are encouraging and matched state-of-the-art

Future work

- Deal with large gaps within binding site patterns
- Improve grammar structure induction
- Speed up convergence of evolution process
- Perform more tests, especially on meta-patterns

References

- [1] Y. Sakakibara, IEEE TPAMI, 7:1051-1062, 2005
- [2] Y. Sakakibara et al., Nucleic Acids Res., 22:5112-20, 1994
- [3] S. Kawashima et al., Nucleic Acids Res., 28:374, 2000
- [4] A. Golovin et al., PROTEINS, 58:190-9, 2005

*Zinc finger meta-pattern is based on 7 PROSITE patterns:

C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H
C-x-H-x-[LIVMFY]-C-x(2)-C-[LIVMYA]
C-x-[DE]-C-x(3)-[LIVMF]-x(1,2)-D-x(2)-L-x(3)-F-x(4)-C-x(2)-C
[FY]-C-x-[DEKSTG]-C-[GNK]-[DNSA]-[LIVMHG]-[LIVM]-x(8,14)-C-x(1,2)-C
C-[DESN]-x([CTS]-x(3)-x(3)-[RK]-x(4)-P-x(4)-[CSLAT]-x(2)-[CAYF]
C-x(2)-C-x(3,5)-[STACD]-x(4)-C-x-[LVFQ]-C-x(4)-[RD]-[NDS]
W-x-C-x(2,4)-C-x(3)-N-x(6)-C-x(2)-C